Solving Linear Systems Faster

A new method that exploits parallel computations promises to have a real impact on practical problems

Two mathematicians recently developed a new method for solving large systems of linear equations on computers—a result that should have immediate applications in numerous fields of science including weather forecasting and economic modeling. In the process, the two resolved a 50-year-old problem. Ronald Rivest of the Massachusetts Institute of Technology describes the new work as "really significant. It promises tremendous speedups and seems likely to have a large practical impact." In addition, Rivest notes, the method is theoretically interesting. "It's a real nice result," he concludes.

The method was developed by Victor Pan of the State University of New York at Albany and John Reif of Harvard University. The two met at a conference several months ago and began talking about ideas to solve linear systems more efficiently. Then Pan visited Reif for a weekend and, to their own surprise, they came upon their method. "It was very exciting," Reif recalls.

Linear systems, which are sets of n linear equations and n variables are among the most fundamental problems in large-scale computing. They are extremely difficult to solve, not because the mathematics is hard but because they can involve thousands of equations in thousands of variables, all of which must be solved by finding values for the variables that satisfy all the linear equations simultaneously. The computations, in short, are onerous. In fact, atmospheric scientists sometimes joke that it takes 3 days of computations to solve the equations to tell you accurately what tomorrow's weather will be.

Mathematicians have developed two different ways to solve these systems of equations. One is to use a direct method such as Gaussian elimination, a procedure that dates back to at least the 1850's and the German mathematician Karl F. Gauss. The idea is to eliminate each variable in a separate stage of the procedure. "It is commonly used and it has the advantage that you get an exact solution," says Reif. "But it has the disadvantage of being inherently sequential," meaning that each stage of the procedure must be performed in order—you cannot speed up the process by doing several stages at the same time.

In addition, when computer scientists try to program the Gaussian elimination method on a computer, they run into problems of stability. Reif explains: "There is a question of how many bits you need to represent a given value to get a good solution. Gaussian elimination uses rational numbers but computers only store rational numbers to some degree of precision. A stable method would give the solution within the degree of precision of the computer but the Gaussian elimination method is potentially unstable since even a small error in precision can snowball, resulting in considerably larger errors in the output." Usually, computer scientists cope with this problem by approximating the rational numbers in the problem and using the Gaussian elimination method to get an approximate solution to their problem. Then they go on to the second method of solving linear equations to refine that solution and make it as exact as they require.

> Linear systems can involve thousands of equations in thousands of variables.

The second type of method for solving linear systems consists of a group of techniques known as iterative methods. These techniques have the advantages of being stable, efficient, and amenable to parallel processing, meaning that many steps of the problem can be worked simultaneously by computers running in parallel. They are not inherently sequential. These methods solve a linear system by inverting an associated $n \times n$ matrix. Once the inverse is known, the linear system can easily be solved by a single matrix multiplication.

But the disadvantage of these methods is that they do not always converge to the solution of the problem. Says Reif, "You have to start with an approximate matrix inverse. Whether the method converges depends on how good that initial approximation is."

So the usual practice is to use Gaussian elimination to get an approximate matrix inverse and then to use an iterative method to finish off the problem. It would be easier to just get an approximate matrix inverse some other way and then go on to use an iterative method directly. However, no one knew how to do that in an efficient way. But people did try.

The first to publish an iterative method that would directly solve problems if an approximate inverse were known was the German mathematician G. Schultz, whose paper appeared in 1933. He showed that his method would give a solution to the problem in approximately $\log n$ stages. In contrast, other iterative methods take about n stages.

Yet, perhaps because no one had a good way to get an approximate matrix inverse, Schultz's work was not well known. Every 10 years or so it was rediscovered. Most recently, Pan and Reif, like all the rest, rediscovered Schultz's method. But they were determined to go on and find an efficient way to get an approximate matrix inverse, which would mean solving a problem that had been stumping researchers for 50 years.

The best previously discovered method for finding matrix inverses was discovered in 1976 by L. Csanky of the University of California at Berkeley. Csanky's method took about $(\log n)^2$ steps but it was unstable and could not be put on a computer because it required an impractical number of processors to do the computations. Other researchers have since managed to decrease the number of processors needed, but the method still is unstable. The method that Pan and Reif developed requires only as many processors as are needed to multiply two $n \times n$ matrices together in $\log n$ steps. This number is now about n^3 but, in theory, it can be reduced to about $n^{2.5}$. In addition, Reif remarks, "The method succeeds in all practical cases."

Reif explains that his improvement in the number of processors required puts him within reach of the theoretically optimal time for finding a matrix inverse. "We're within a log factor of optimal," he says. "That's a considerable improvement over previous bounds."

In addition, Pan and Reif made their method much more efficient for certain commonly occurring problems in which

the matrices have a lot of zeros—"sparse systems." These systems occur, for example, in weather forcasting and in the design of airplane wings. When a linear system is appropriately sparse, Reif and Pan's method uses far fewer processors than it does for more dense matrices. Now, says Reif, "In theory, you can get an answer at least an order of magnitude faster."

Yet, according to Reif, most parallel processors in use today have far fewer than 1000 processors and so the speed-up with the new method is currently less dramatic than it could be, although it is still substantial. But already there are a few systems with huge numbers of processors. For example, the Thinking Machine Inc.'s Connection Machine in Cambridge, Massachusetts, has 16,000 processors and the company is now developing a network with 64,000 processors. Several other companies, including IBM, claim that they intend to build networks with more than a thousand processors within the next few years. So, with the new parallel processors and the new algorithm, the process of finding solutions to huge linear systems should be much quicker. Weather forecasting equations, Reif notes, should be much easier to solve. In short, says Rivest, "the possibilities look very exciting."

—GINA KOLATA

Additional Reading

1. V. Pan and J. Reif, "Efficient parallel solution of linear systems," in *Proceedings of the 17th Annual ACM Symposium on the Theory of Computing*, Providence, R.I., May 1985, pp. 143–152.

Something Strange from Cygnus X-3

At least two proton decay experiments have now detected particle showers that seem to be triggered by emissions from the galactic x-ray source Cygnus X-3. However, the emissions are baffling: known elementary particles, such as photons or neutrinos, can be ruled out. Nor is there an obvious candidate among the supersymmetric and grand unified particles concocted by the theorists. If real, the Cygnus X-3 particle would have to be something new.

Cygnus X-3 itself is an x-ray binary system with a period of 4.79 hours, lying some 30,000 light-years from Earth. Essentially it consists of a compact object, probably a neutron star, pulling a stream of gas from a more or less normal companion star; in the process the gas is heated sufficiently to emit the x-rays observed. In fact, Cygnus X-3 is probably the most powerful source of high-energy photons in the galaxy. It is also well situated in the northern sky for observation by many of the proton decay experiments. The first indications came about 2 years ago, when showers of muons from the general direction of Cygnus X-3 were seen in a prototype detector operated in Minnesota's Soudan iron mine by physicists from the University of Minnesota and the Argonne National Laboratory.

As it happens, when the Soudan group submitted their results for publication, one of the reviewers was John Learned of the University of Hawaii, a member of the team that operates the giant Irvine-Michigan-Brookhaven (IMB) detector in the Morton Thiokol salt mine near Cleveland. Following Soudan's lead, Learned started analyzing the IMB muon events with particular attention to Cygnus X-3; by Christmastime 1984, he and his students had found suggestions of a signal that matched both the 4.79-hour periodicity of Cygnus X-3 and its proper phase.

Learned accordingly passed word back to the Soudan group and to the proton decay community at large. The Soudan physicists have now reanalyzed their data and confirmed the result: out of 874,000 muon events, 1200 come from the general direction of Cygnus, and an excess of 80 show the 4.79-hour periodicity (*1*). "It's like picking out a lighthouse," says Minnesota's Marvin L. Marshak. Similar results have also been reported from the European NUSEX detector in the Mont Blanc tunnel (*2*).

There remains the question of what is causing the muon tracks. Since muons are unstable and short-lived, they are presumably produced by some kind of primary particle from Cygnus X-3 interacting with the earth's atmosphere or with the rock around the detectors. (The detectors experience an enormous background flux of muons produced by cosmic rays in exactly this way.)

The fact that the periodicity is detectable over a distance of 30,000 light-years means that all the primary particles have to be moving at the same speed, the speed of light; otherwise some would lag behind the others and the signal would be washed out. The fact that the primaries still show some directionality means that they must be electrically neutral; otherwise the galactic magnetic field would have deflected and randomized them.

The only known particles that fit those two criteria are neutrinos, photons, and ultrahigh-energy neutrons. However, neutrons can be ruled out because they themselves are unstable, with a 15.3-minute half-life. To survive the trip they would need an energy in excess of 10^{18} electron volts. Yet the flux of all known cosmic rays above that energy would produce only about one event per year in the Soudan detector.

Neutrinos can be ruled out by the zenith angle effect: the signal tends to die away as Cygnus X-3 approaches the horizon, as if the primaries were being absorbed by the atmosphere or the surrounding rock. Neutrinos are perfectly capable of traversing the whole earth and would produce an isotropic distribution of muons.

And finally, photons can be ruled out because they simply do not produce enough muons. Barring some previously unsuspected interaction mechanism, calculations show that the known flux of high-energy photons from Cygnus X-3 fails to produce enough muon showers by a factor of 300.

"If the results are right, the deficiencies [with known particles] are *gross*," says Learned. "There's no way the theorists can wiggle out with a factor of 2 here or there. The only question is, Are the experiments correct?"

Indeed, there is ample reason to be cautious: the IMB collaboration in particular has been looking at additional data using two independent methods of analysis and has so far been unable to verify Learned's signal. As Learned himself paraphrases the group's official stance, "Whatever it is we do or don't see, it isn't neutrinos."

—M. MITCHELL WALDROP

References

1. M. L. Marshak *et al.*, *Phys. Rev. Lett.* **54**, 2079 (1985).
2. G. Battistoni *et al.*, submitted to *Phys. Lett. B*.